John von Neumann Institute for Computing

Protein Structure Prediction Using Coarse Grain Force Fields

N. Mahmood, A. Torda

published in

From Computational Biophysics to Systems Biology (CBSB08),
Proceedings of the NIC Workshop 2008,
Ulrich H. E. Hansmann, Jan H. Meinke, Sandipan Mohanty,
Walter Nadler, Olav Zimmermann (Editors),
John von Neumann Institute for Computing, Jülich,
NIC Series, Vol. **40**, ISBN 978-3-9810843-6-8, pp. 309-312, 2008.

http://www.fz-juelich.de/nic-series/volume40

Protein Structure Prediction Using Coarse Grain Force Fields

Nasir Mahmood and Andrew Torda

Center for Bioinformatics, University of Hamburg,
Bundesstrasse 43, D-20146 Hamburg, Germany
E-mail: mahmood@zbh.uni-hamburg.de

In ab initio or de novo protein modelling, one tries to build 3D protein models from scratch rather than modelling them on to known structures. Our method is based on special purpose low resolution force fields. They are rather different to most approaches by not taking into account any strict physical model. They are statistical, but there is no assumption of Boltzmann statistics. In a Monte Carlo simulation, the acceptance criterion can be directly based on the calculated probabilities. Although we have not performed proper benchmark, the scoring function works reasonably well to predict 3D models of smaller proteins from their sequences.

1 Introduction

Protein structure prediction is one of the classic problems from computational chemistry or molecular structural biology. Essentially, one would like to be able to go from the sequence of a protein (easily obtained) to the structure (expensive and often difficult to obtain experimentally). Our interest has been in devising new purely probabilistic score functions. They make no use of Boltzmann statistics, but instead rely on a mixture of Bayesian probabilities based on normal and discrete distributions. This has an interesting consequence if one works with a method such as Monte Carlo, one can base the acceptance criterion directly on the calculated probabilities without assuming a Boltzmann distribution. Monte Carlo simulations, in their various forms, have been described by reputable scientists as the path to the simulator's graveyard. This poses the question as to why a rational simulator would venture further into this field. There are two aspects to this problem: 1) the score or quasi-energy function and 2) the search method. The score function may be energy-like or purely statistical and the search method is used to explore the conformational space. The score function and search method are often coupled together and search method is driven by score function to get to native like structures.

2 Method

2.1 Score Function

Unlike most Monte Carlo methods we do not use an energy or score, but calculate probabilities (or ratio of probabilities) directly:

$$Probability\,ratio = \frac{P(X_N)}{P(X_O)} \tag{1}$$

$P(X_O)$ and $P(X_N)$ are probabilities of old and updated conformations respectively.

$$P(X|\vec{V},T,S) = \prod_i [\sum_j (\pi_j \prod_k P(X_{ik}|X_i \epsilon C_j, \vec{V}_{jk}, T_{jk}, S))] \quad (2)$$

$X = X_1, ..., X_I$	the set data instances X_i (fragments)
$X_i = \vec{X_{i1}}, ..., X_{iK}$	attribute vectors X_{ik} describing X_i
i	observation number, $i = 1, ..., I$
j	class number, $j = 1, ..., J$
k	attribute number, $k = 1, ..., K$
c	inter-class probabilities and parameters
S	the set of possible probability density functions (p.d.f.) covering $\vec{V}, T$
$T = T_c, T_1, ..., T_J$	the exact functional form of each p.d.f.
$\vec{V} = \vec{V_c}, \vec{V_1}, ..., \vec{V_J}$	the set of parameters instantiating p.d.f.
π_j	class mixture probability, $\vec{V_c} = 1, ..., J$

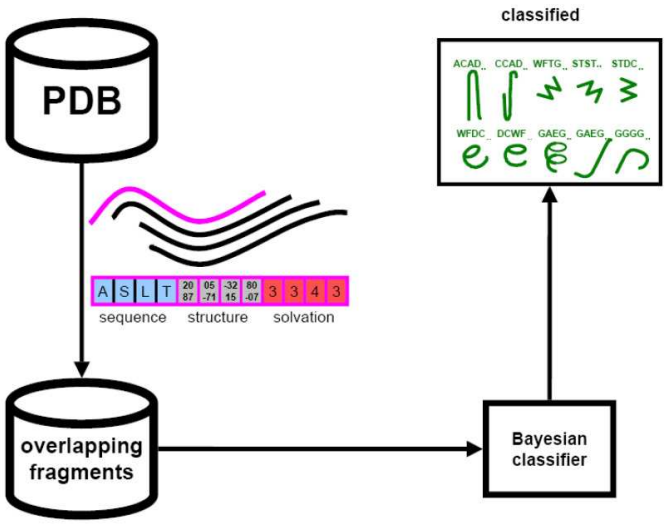

Figure 1. Bayesian classification: overlapping fragments generated from existing structures classified into a number classes by Bayesian classifier. Each fragment is represented by its sequence, structure & solvation.

Our score function is purely probabilistic and relies on mixture of Bayesian probabilities by combing sequence, structure and solvation. The statistical models: multi-way Bernoulli, bivariate Gaussian and simple Gaussian were used to model sequence, structure and solvation respectively (see figure 1).

2.2 Search Method

We are using simulated annealing Monte Carlo as a search method to find the most probable structural arrangement of a given amino acid sequence.The search method makes two kinds of moves: 1) biased moves made by drawing a fragment from a fragment library generated from existing protein structures and 2) completely unbiased moves. Internally, the score function is based on dihedral angles, Cartesian coordinates and sequence description, so there is some computational work involved in moving between representations. The acceptance criterion depends solely upon the probability ratio (equation 1) calculated from the probabilities of the new and old structures.

3 Results

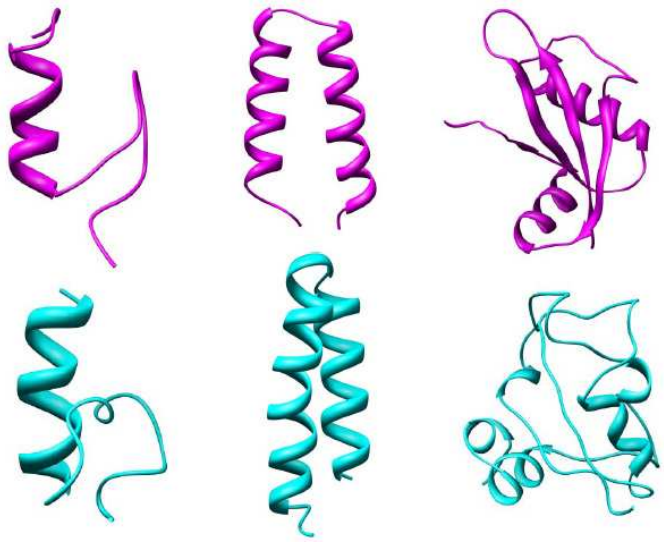

Figure 2. Top row: native structures 1fsv, 2hep and 2hfq from left to right, bottom: respective predicted models.

4 Conclusion

The current implementation seems to have a rather good representation of local interactions and works surprisingly well for small proteins. The score function has also been integrated with our existing protein threading machinery to be used for CASP8 competition. We are now working on incorporating simple solvation and hydrogen bond effects into the initial probability calculations to better account for long range interactions.